BIOINFORMATICS ORIGINAL PAPER

Vol. 26 no. 14 2010, pages 1731–1737
doi:10.1093/bioinformatics/btq272

Genetics and population analysis

Advance Access publication May 25, 2010

A quality control algorithm for filtering SNPs in genome-wide

Monnat Pongpanich[1], Patrick F. Sullivan[2] and Jung-Ying Tzeng[1,3,*]

[1]Bioinformatics Research Center, North Carolina State University, Campus Box 7566, Raleigh, NC 27695-7566, [2]Department of Genetics, University of North Carolina at Chapel Hill, Campus Box 7264, Chapel Hill, NC 27599 and [3]Department of Statistics, North Carolina State University, Campus Box 7566, Raleigh, NC 27695-7566, USA

Associate Editor: Martin Bishop

ABSTRACT

Motivation: The quality control (QC) filtering of single nucleotide polymorphisms (SNPs) is an important step in genome-wide association studies to minimize potential false findings. SNP QC commonly uses expert-guided filters based on QC variables [e.g. Hardy–Weinberg equilibrium, missing proportion (MSP) and minor allele frequency (MAF)] to remove SNPs with insufficient genotyping quality. The rationale of the expert filters is sensible and concrete, but its implementation requires arbitrary thresholds and does not jointly consider all QC features.

Results: We propose an algorithm that is based on principal component analysis and clustering analysis to identify low-quality SNPs. The method minimizes the use of arbitrary cutoff values, allows a collective consideration of the QC features and provides conditional thresholds contingent on other QC variables (e.g. different MSP thresholds for different MAFs). We apply our method to the seven studies from the Wellcome Trust Case Control Consortium and the major depressive disorder study from the Genetic Association Information Network. We measured the performance of our method compared to the expert filters based on the following criteria: (i) percentage of SNPs excluded due to low quality; (ii) inflation factor of the test statistics (λ); (iii) number of false associations found in the filtered dataset; and (iv) number of true associations missed in the filtered dataset. The results suggest that with the same or fewer SNPs excluded, the proposed algorithm tends to give a similar or lower value of λ, a reduced number of false associations, and retains all true associations.

Availability: The algorithm is available at http://www4.stat.ncsu.edu/~jytzeng/software.php

Contact: jytzeng@stat.ncsu.edu

Supplementary information: Supplementary data are available at *Bioinformatics* online.

Received on February 27, 2010; revised on April 26, 2010; accepted on May 20, 2010

1 INTRODUCTION

Genome-wide association studies (GWAS) have been shown to be a powerful and successful strategy in identifying genetic variants that influence common and complex diseases. Prior to the advent of GWAS in 2005, there were only a few robust, replicated associations identified, such as *NOD2* for Crohn's disease (CD; Hugot *et al.*, 2001), and PPARG, KCNJ11 and CAPN10 for Type 2 diabetes (T2D) mellitus (McCarthy, 2004). With GWAS, there are now more than 30 loci identified for CD and almost 20 loci for T2D (Barrett *et al.*, 2008; Zeggini *et al.*, 2008). To date (April 2010), there are over 545 published studies reporting genetic variants responsible for more than 340 common diseases (Hindorff *et al.*, 2009; http://www.genome.gov/gwastudies).

GWAS interrogate millions of single nucleotide polymorphisms (SNPs), and the large-scale genotype calling (which translates probe hybridization intensities into actual genotypes) must fully resort to automated clustering procedures (Plagnol *et al.*, 2007; Teo, 2008). Ideally, SNP genotyping yields three clusters of signals, and a subject's genotype can be assigned according to cluster membership (Ziegler *et al.*, 2008). In reality, the clustering methods are unavoidably prone to error, as imperfect clusters of signal clouds can arise due to experimental variation, DNA quality and non-specific hybridization issues (Anney *et al.*, 2008; Clayton *et al.*, 2005). Common error patterns include missing calls for SNPs with overlapping genotype clusters (Anney *et al.*, 2008), homozygote–heterozygote miscalls (Teo *et al.*, 2007), false homozygote calls in heterozygous individuals due to allelic dropout (Pompanon *et al.*, 2005), and erroneous assessment of monomorphic SNPs as polymorphic (Pettersson *et al.*, 2008).

SNP quality control (QC) is commonly safeguarded by 'supervised' (i.e. expert-guided) filters to exclude low-quality SNPs. The 'supervised' expert filters aim to remove SNPs that fall into the extremes of QC variables including Hardy–Weinberg equilibrium (HWE), missing proportion (MSP) and minor allele frequency (MAF). The rationale is clear: extreme deviation from HWE is typically used to identify gross genotyping error (Teo *et al.*, 2007); a high MSP indicates poor genotype probe performance and low genotyping accuracy (Neale and Purcell, 2008; WTCCC, 2007); SNPs with low MAF are more prone to error, as fewer samples would be within a genotype cluster and most clustering-based calling algorithms do not perform well with rare alleles (Neale and Purcell, 2008; Teo, 2008). However, the implementation of expert filters tends to require arbitrary determination of cutoff values for the QC variables, and does not jointly consider all QC features. For example, in GAIN studies, the minimum SNP genotyping quality standards are HWE P-value >0.00033, average MSP $< 3\%$,

*To whom correspondence should be addressed.

MSP maximum <10% and quality score and MAF greater than a pre-determined minimum level, which varies from study to study (GAIN Collaborative Research Group, 2007). For GWAS conducted by the Wellcome Trust Case Control Consortium (WTCCC), the criteria for retaining a SNP are: HWE P-value $\geq 5.7 \times 10^{-7}$, MSP $\leq 5\%$ if MAF $\geq 5\%$, MSP $\leq 1\%$ if MAF $< 5\%$ and MAF > 0.01 (WTCCC, 2007). Sladek *et al*. (2007) included SNPs when the HWE P-value >0.001, MSP $\leq 5\%$ and MAF > 0.01. Unoki *et al*. (2008) included SNPs when the HWE P-value $\geq 10^{-6}$ and MSP $\leq 10\%$.

Statistical methods have also been developed to identify, assess or incorporate genotyping errors in association studies (Gordon *et al*., 2001; Gordon and Ott, 2001; Hao and Wang, 2004; Rice and Holmans, 2003). Recently, Plagnol *et al*. (2007) introduced a calling algorithm to minimize the biases that occur when case and control DNA samples are from different sources and processed in different laboratories. Miyagawa *et al*. (2008) investigated appropriate cutoff values for each of the QC variables (MSP, MAF, HWE and confidence score of genotype calls) by dividing and reshuffling healthy samples. Teo *et al*. (2008) assessed the stability of the assigned genotypes by introducing white noise to the fluorescent intensities of each subject and evaluating the agreement between the calls made with the noise-perturbed and original intensities. Finally, for family-based studies, Fardo *et al*. (2009) developed a transmission test to measure the genotyping error rate of each proband.

In this work, we take the rationale of the expert filters and propose an 'unsupervised' (i.e. algorithm-determined) filter to detect low-quality SNPs. Like 'supervised' expert filters, our filter also aims to identify QC outliers. Furthermore, our filter automates the QC threshold determination based on all QC features, and gives conditional cutoffs contingent on the values of other QC variables (e.g. different MSP thresholds for different MAFs). The algorithm is based on the premise that the majority of SNPs have sufficient genotyping quality with QC variable values in certain directions (e.g. low MSP and non-low MAF). SNPs with QC values deviating from the majority are considered outliers and are then labeled as problematic SNPs. The algorithm first performs principal component analysis (PCA) on the QC variables with an aim to separate good SNPs from problematic SNPs on a two-dimensional plane. It then uses Density Based Spatial Clustering of Applications with Noise (DBSCAN; Ester *et al*., 1996) to identify the boundaries of good SNPs and define QC thresholds. We evaluate the performance of the proposed algorithm and demonstrate its utility using the seven WTCCC datasets (WTCCC, 2007) and the major depressive disorder (MDD) dataset from Genetic Association Information Network (GAIN) studies (Sullivan *et al*., 2009).

2 METHODS

2.1 The proposed QC algorithm

We begin with a SNP dataset that has been cleaned using the criteria of quality score and HWE. That is, if an SNP does not reach the desired level of quality score, the genotyping result is specified as 'missing'. In addition, SNPs that show severe HWE violations in the control group (i.e. the P-value of the HWE test is smaller than a threshold appropriate for multiple testing) are excluded from the dataset. Quality score and HWE are used to pre-clean the dataset because they have clear definitions for good SNPs. While deviation from HWE has relatively low sensitivity in testing for genotyping error (Cox and Kraft, 2006), it has been shown that severe genotyping errors often do cause extreme HWE deviations (Teo *et al*., 2007).

With this pre-cleaned dataset, our algorithm aims to identify good-quality SNPs based on two basic QC features, MSP and MAF. Specifically, we consider six QC variables including MSP in case samples (denoted by MSPcs), MSP in control samples (MSPcn), MSP in the combined case-control samples (MSPall), logMAF in the combined samples (logMAFall), the ratio of MSPcs to MAFcs and the ratio of MSPcn to MAFcn. MAF is considered on the log scale to ensure a more careful QC examination with a low MAF than a high MAF. The interaction term between MSP and MAF is designed to allow for an adaptive MSP threshold with different MAF values, and is defined as MSP $\times$ (1/MAF). The adaptive thresholds ensure that SNPs with smaller MAF have a more stringent MSP threshold, as missing genotypes have a larger impact on frequency when occurring in low MAF than in high MAF. We use the ratio rather than the product of MSP and MAF, so that different low-quality features (e.g. high MSP and low MAF) will be retained rather than being cancelled out in the interaction terms. A higher interaction value indicates lower quality.

There are two main steps involved in the proposed QC algorithm: (i) using PCA to separate the good SNPs from the bad SNPs based on the QC features on a two-dimensional plane; and (ii) using DBSCAN (Ester *et al*., 1996) to identify the boundaries of good SNPs on the plane. The PCA is performed on the six QC variables to separate good SNPs from bad ones on the plane of the first two principal components (PC1 versus PC2), which usually account for about >80% of the variation in the original QC variables. The use of PCA facilitates the task of modeling all of these QC variables that can be highly correlated. It also projects good SNPs into a concentrated corner on the plane and spreads out bad SNPs in opposite directions along the axes of the original QC variables (e.g. see the biplots shown in Figs 1a and 2a, and the expert SNP classification in Figs 1b and 2b). Different studies may result in different patterns of PC biplots, but the key common feature is that good SNPs are pushed toward a certain corner that represents desirable QC values: low MSP, high MAF and low MSP to MAF ratio.

Given the PCA plots, we use DBSCAN (Ester *et al*., 1996) to define the boundaries of the good SNPs. DBSCAN is a density-based clustering algorithm, it performs efficiently on large-scale datasets, and most importantly, it can find clusters of arbitrary shape. Given a data space, it defines regions of high-density points as *clusters* and classifies regions of low-density points as *noises* (i.e. a noise is a point that does not belong to any clusters). DBSCAN requires that for each point in a cluster, there are at least a minimum number, K, of points in the neighborhood of a given radius r of the target point. Ester *et al*. recommended setting K to four, and to determine r from the data via the following steps. First, calculate the distance of each target point to its K-th nearest point. Next, plot the sorted K-th nearest neighbor (NN) distance (which is referred to as the sorted K-th NN graph). Finally, set r to the Y-axis value where a sharp jump occurs. We follow the suggestion of using $K = 4$. Instead of eyeballing the value for r as suggested by the original work, we solve for r by fitting a change point model as described in the Appendix A1. The r value determined by the change-point method should be viewed as an initial value, and should be further fine tuned until certain criteria are fulfilled. For example, tune r until the resulting 'good' SNPs yield a desirable λ value (i.e. the inflation factor of the test statistics; Devlin and Roeder, 1999), until maximum MSP is smaller than a desirable level, or until a certain percentage of SNPs are removed. With the suitable r value, we then use the largest cluster identified by DBSCAN to define the boundaries for good SNPs (e.g. see the borders of the blue area in Figs 1c and 2c). The boundaries represent meaningful thresholds with respect to the original QC variables. In the final output of the algorithm, an SNP is labeled as 'good' if (i) it is in the largest cluster, or (ii) it passes the identified thresholds of all QC features even if it is not in the largest cluster. Criterion (ii) ensures the monotonicity of the thresholding. That is, any SNPs located in the 'good SNP corner' (i.e. high logMAF, low MSP and low MSP/MAF) will be included even if they are not dense enough to be included in a cluster.

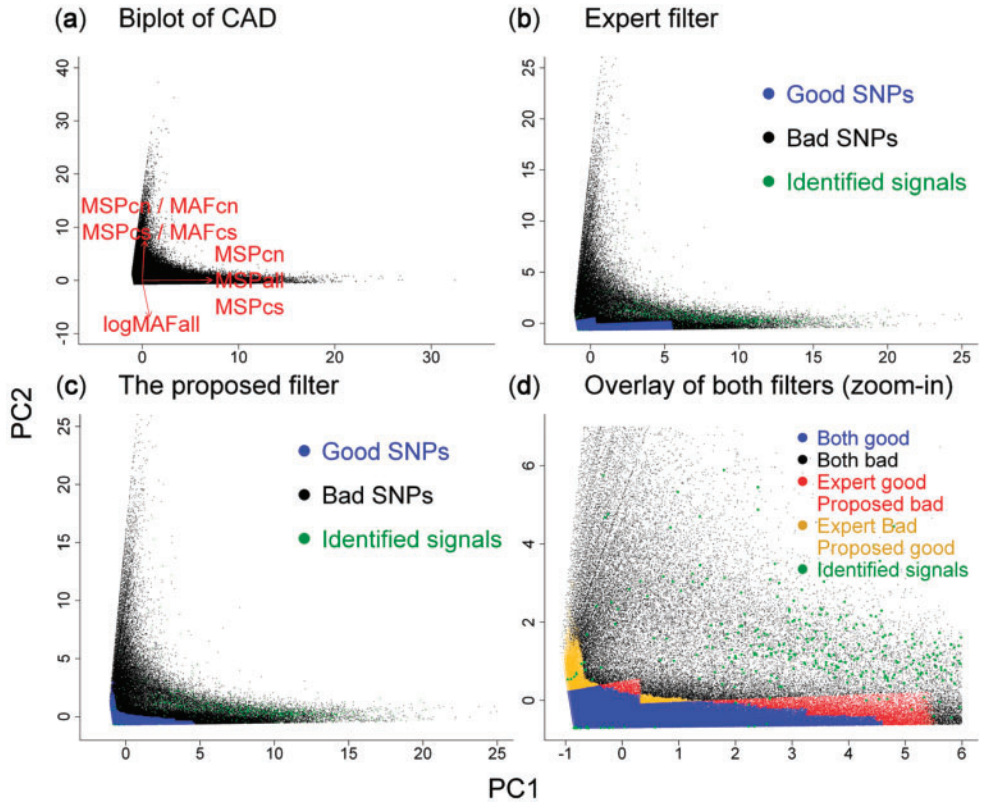

Fig. 1. Projections of SNPs on the two-dimensional PC plane for the WTCCC CAD study. (**a**) PCA biplot, where the red arrows represent directions of the original variables. (**b**) SNP classification results from the expert filter. (**c**) SNP classification results from the proposed filter. (**d**) Overlay of good SNP boundaries defined by both filters. In (b), (c) and (d), the identified signals (green dots) are those SNPs with association test *P*-values $<5\times10^{-7}$.

2.2 Performance evaluations using real datasets

The performance of the proposed QC algorithm was evaluated using the seven GWAS studies conducted by WTCCC, including bipolar disorder (BD), coronary artery disease (CAD), CD, hypertension (HT), rheumatoid arthritis (RA), Type 1 diabetes (T1D), and T2D (WTCCC, 2007). In addition, we also assess the algorithm using the MDD dataset from GAIN studies (Sullivan *et al*., 2009). In each WTCCC GWAS, there were 490 032 SNPs genotyped on chromosomes 1 to 22 from 2000 cases and 3000 common controls, which included 1500 from the 1958 British Birth Cohort (58C) and another 1500 from blood donors recruited by UK Blood Services. We excluded unreliable individuals as defined in the original studies: poor sample call rate (<97%), extreme overall heterozygosity (>30% or <23%) and high genome-wide IBD values (>0.86), and obtained on average 1887 cases and 2974 controls. We then removed those SNPs with HWE *P*-value $<5.7\times10^{-7}$ (WTCCC, 2007) and were left with 474 657 SNPs for the QC evaluations. The MDD study contained 556 131 SNPs genotyped on chromosomes 1 to 22 from 1738 MDD cases and 1802 controls. In this dataset, all unreliable samples (e.g. poor sample call rate, extreme heterozygosity, high relatedness and ancestral outliers) have been excluded using the steps described in Sullivan *et al*. (2009). We removed SNPs with HWE *P*-value $<5.7\times10^{-7}$ and performed the QC analysis on the remaining 526 740 SNPs.

The results of the proposed QC algorithm are compared with the expert filter defined in WTCCC (WTCCC, 2007), which removed SNPs with MSP > 5% if MAF ≥ 5%, MSP > 1% if MAF < 5% or SNPs with MAF ≤ 1%. The performances are assessed based on the following four criteria: (i) percentage of SNPs excluded due to low quality; (ii) inflation factor of the substructure-adjusted test statistics λ; (iii) number of false associations found in the filtered dataset [referred to as false positives (FP)]; and (iv) number of true associations missed in the filtered dataset [referred to as true positives (TP)]. For (ii), the inflation factor λ is calculated as the median of the observed test statistics of association divided by the median of $\chi^2_{(1)}$ distribution (i.e. 0.456) (Devlin and Roeder, 1999). For the WTCCC datasets, the association statistics were calculated using a stratified Cochran–Mantel–Haenszel test in PLINK (Purcell *et al*., 2007) to adjust for population substructure. For the MDD dataset, the trend test statistics were used because the samples are ancestrally homogeneous (Sullivan *et al*. 2009). For (iii), an FP is defined as a significant signal found in the data analyses but not confirmed in the literature (i.e. neither in PubMed database nor the published GWAS catalog at www.genome.gov/gwastudies). For (iv), a TP is defined as a significant signal found in the data analyses and also confirmed in the literature (either in PubMed or the published GWAS catalog). A *P*-value threshold of 5×10^{-7} (following the WTCCC paper) is used to define significance. However, because there were no literature-confirmed signals that survived the 5×10^{-7} threshold for BD, HT and MDD, we used a less stringent threshold of 10^{-5} for the *P*-value in our analysis for these three diseases. A threshold of 10^{-5} is considered to be a moderate association in the WTCCC studies (WTCCC, 2007).

2.3 Implementation

We have implemented a command-line-based software package to perform the proposed QC algorithm. The software runs PCA using the 'prcomp' function in R and runs DBSCAN using C++ code written by us for speed

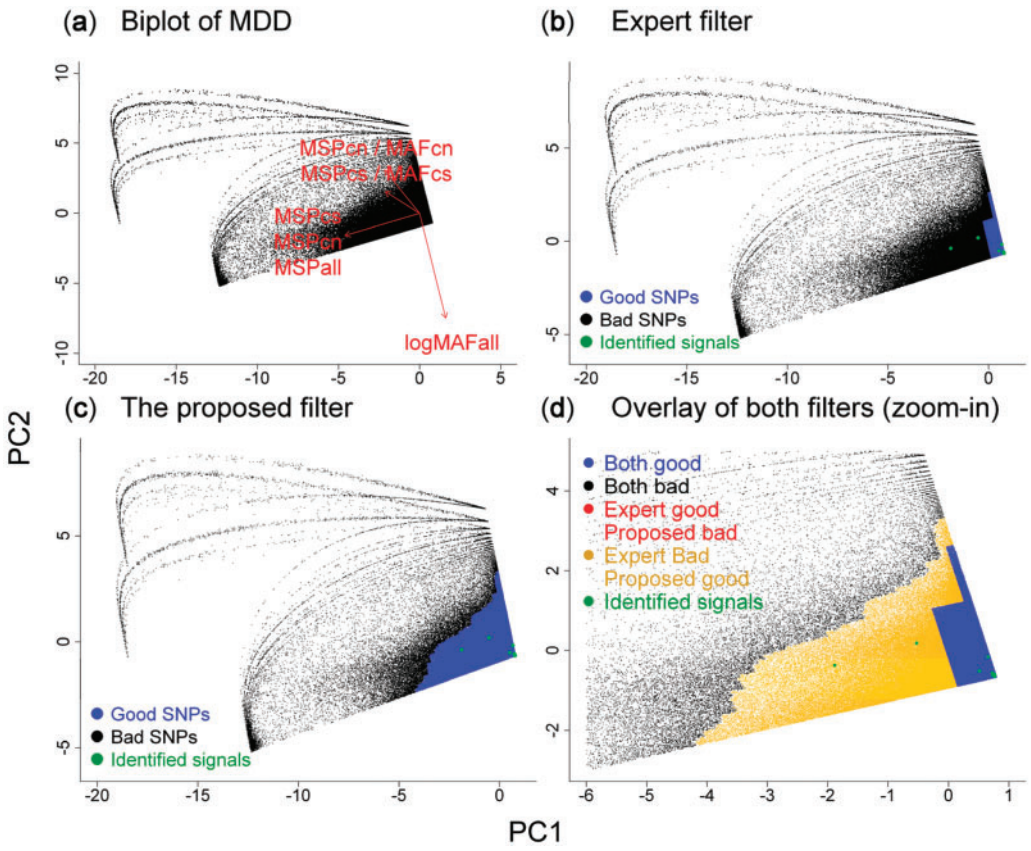

Fig. 2. Projections of SNPs on the two-dimensional PC plane for the GAIN MDD study. (**a**) PCA biplot, where the red arrows represent directions of the original variables. (**b**) SNP classification results from the expert filter. (**c**) SNP classification results from the proposed filter. (**d**) Overlay of good SNP boundaries defined by both filters. Note in (d) there are no red dots because the good SNP region by the proposed method is a superset of the expert good region. In (b), (c) and (d), the identified signals (green dots) are those SNPs with association test *P*-values $< 10^{-5}$.

improvement. The software and instructions are available for download from the corresponding author's website.

3 RESULTS

Figure 3 shows the results of our method and the WTCCC expert filter based on the four criteria. The specific numerical results are given in the Supplementary Tables 1 and 2. To illustrate, we report the results using the r value obtained by the change-point model for all eight diseases regardless of whether a further fine-tuning of r was carried out. Overall, the algorithm with change-point r removed from 2.8% to 14.6% fewer SNPs than the expert filters, and yet had either smaller or comparable λ values, contained fewer or comparable FPs, and retained the same TPs (which were all the TPs in the genotyped SNPs). The maximum MSP retained in the datasets ranged from 4.44% to 5.55% for the WTCCC datasets and was 37.63% for MDD.

Carefully examining Supplementary Tables 1 and 2, we saw that there were three diseases where the performance with the initial change-point r value was not as good as expert filters in some of the criteria: BD (having a larger $\lambda = 1.123$ than the 1.122 of the expert filter), RA (having a larger $\lambda = 1.083$ than 1.052 of expert and including four more FPs) and MDD (having two more FPs than expert). Using BD as an example, with the change-point r value, our filter removed 13.2% of the SNPs (versus 18.6% of expert), and the resulting 'good' SNPs had a maximum MSP of 4.91%, a not-small λ value of 1.123 (versus 1.122 of expert), 19 FPs (versus 27 of expert, out of 912 unfiltered FPs) and retained all 6 TPs (same as expert). In practice, the algorithm should be continued by adjusting r until a desirable λ is reached. However, for comparison purposes, we instead fine-tuned r to a smaller value so that the two filters removed about the same proportion of SNPs. With a similar removal rate (18.2% versus 18.6% of expert), our algorithm gave a slightly smaller λ (1.114 versus 1.122) and kept fewer FPs (18 versus 27). The results of the TPs remained unchanged.

In RA, we removed 9.8% of SNPs (versus 18.7% of expert), which resulted in a maximum MSP of 5.55%, a λ of 1.083 (versus 1.052 of expert), 211 FPs (versus 207 of expert, out of 817 unfiltered FPs), and the same number of TPs as the expert filter (6 out of 6 unfiltered TPs). When we removed about the same proportion of SNPs as the expert filter (18.5% versus 18.7%), the remaining good SNPs yielded a slightly smaller λ (1.048 versus 1.052), contained 16 fewer FPs (191 versus 207) and the same number of TPs (6). In MDD, with the change-point r, the algorithm removed much fewer SNPs (5.26% versus 19.89%), yielded comparable λ (1.043 versus 1.044), but kept two more FPs (6 versus 4 out of 6 unfiltered FPs) compared to the

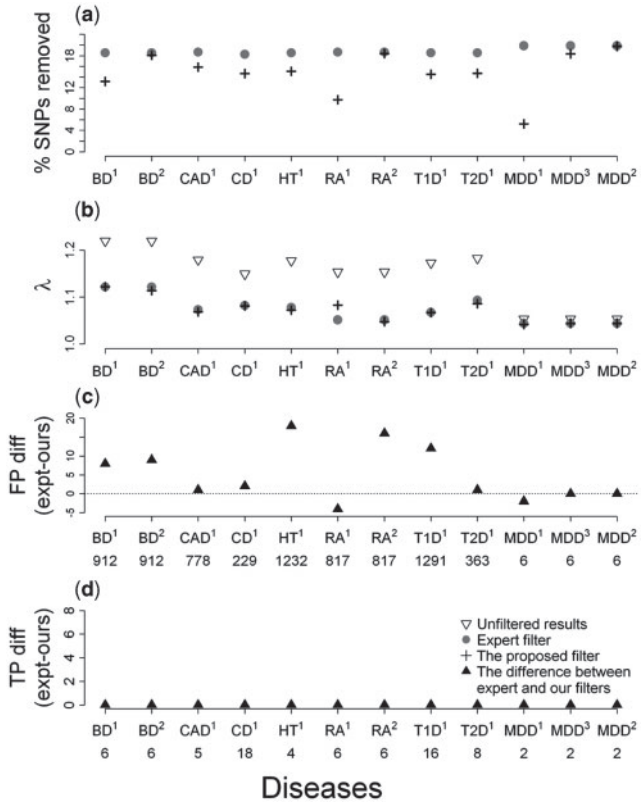

Fig. 3. Performance comparisons of different filters based on the four criteria defined in the text. (**a**) Percentage of SNPs removed. (**b**) λ from substructure-adjusted test statistics. (**c**) Difference (expert-proposed) in the numbers of FPs retained by different filters. The numbers below each disease code on the *X*-axis show the counts of FPs in the unfiltered results. (**d**) Difference (expert-proposed) in the numbers of TPs retained in the dataset by different filters. The numbers below each disease code on the *X*-axis show the counts of TPs in the unfiltered results. On the *X*-axis of each figure, '[1]' indicates the results based on the *r* value determined by the change-point model, '[2]' indicates the results based on the *r* value that makes the proportion of SNPs removed by the proposed filter comparable to the expert filter and '[3]' indicates the results based on the *r* value that makes the maximum MSP in the filtered dataset <10%.

Table 1. Agreement and disagreement in SNP classifications (good SNPs versus bad SNPs) by the proposed filter and the WTCCC expert filter

Diseases	*r*	Agreed		Disagreed	
		Both good[a] (%)	Both bad[b] (%)	Good versus bad[c] (%)	Bad versus good[d] (%)
BD	0.0154[e]	80.33	12.10	6.47	1.10
	0.0110[f]	78.38	15.10	3.47	3.06
CAD	0.0125[e]	79.32	13.92	4.78	1.99
CD	0.0134[e]	79.58	12.57	5.72	2.13
HT	0.0132[e]	79.69	13.37	5.20	1.74
RA	0.0179[e]	80.76	9.27	9.45	0.53
	0.0101[f]	77.05	14.23	4.48	4.24
T1D	0.0133[e]	79.62	12.77	5.80	1.81
T2D	0.0136[e]	79.93	13.24	5.34	1.50
MDD	0.0259[e]	80.11	5.26	14.63	0.00
	0.0072[g]	78.95	17.26	2.63	1.16
	0.0063[f]	78.53	18.24	1.65	1.58

[a]The % of SNPs classified as 'good' by both filters.
[b]The % of SNPs classified as 'bad' by both filters.
[c]The % of SNPs classified as 'good' by the proposed filter but 'bad' by the expert filter.
[d]The % of SNPs classified as 'bad' by the proposed filter but 'good' by the expert filter.
[e]The *r* value is determined by the change-point model.
[f]In these analyses, the values of *r* are chosen to make the proportion of SNPs removed by the proposed filter comparable to that of the WTCCC expert filters.
[g]The *r* value is chosen to make the maximum MSP in the resulting good SNPs <10%.

expert filter. Because the resulting maximum MSP was too large (i.e. 37.6%) when we used the change-point *r*, we tuned *r* by decreasing its value till the maximum MSP was <10% (i.e. 9.7%). The λ and FPs became 1.044 and 4, respectively, which were the same as the expert filters. The λ and FPs stayed unchanged when we continued tuning *r* until we had removed the same proportion of SNPs as the expert filter.

We also categorized the SNPs into four groups according to whether they were included (i.e. labeled as 'good SNP') or excluded (i.e. labeled as 'bad SNP') by our filter and the expert filter (Table 1). For all of the diseases, our algorithm and the expert filter have around 80% agreement in inclusion and around 12% agreement in exclusion on average. The majority of the disagreement between the two filters can be attributed to the use of adaptive thresholds in our filter. To illustrate, we show the boundary of good SNPs from our filter, the expert filter, and the overlay of the two on a two-dimensional PC plane using CAD (Fig. 1d) and MDD (Fig. 2d) (see Supplementary Fig. 1 for other diseases). Instead of the step-like boundary of good SNPs in the expert filter, our filter gives a smoother boundary of good SNPs. Figure 4 further illustrates the disagreements on the axes of MSP versus MAF (instead of PC1 versus PC2) for CAD (see Supplementary Fig. 2 for other diseases). The yellow dots represent SNPs that are labeled as 'good' by our algorithm but 'bad' by expert filters. One group of yellow dots occurred in the extremely low MAF and low MSP range, indicating that our algorithm would keep SNPs of MAF < 0.01 when their MSPs were extremely low. In contrast, the red dots represent the SNPs that are labeled as 'bad' by our algorithm but 'good' by expert filters. The big red area on the right side indicates that our algorithm gives a more stringent MSP criterion for good SNPs than the expert filter (i.e. MSP < 5%): our criteria ranged from MSP < 2% to MSP < 4%, depending on the MAF. Lastly, the two red and yellow triangles in the upper middle area show the impact of the 'smoother' threshold of our algorithm: it has a more stringent MSP threshold when MAF is 0.01–0.025, and a less stringent threshold when MAF is 0.025–0.05.

4 DISCUSSION

Ensuring the quality of genotype data is essential for drawing accurate and replicable conclusions (Donnelly, 2008). In this work, we have introduced a QC algorithm to identify SNPs with low QC features using criteria determined through PCA and DBSCAN. The proposed filter is in essence an 'unsupervised' (i.e. algorithm-determined) version of the 'supervised' expert filter to classify SNPs, and it aims to account for multiple QC variables, provides adaptive cutoff values and automates thresholding decisions. Specifically, we use PCA to jointly model the potentially highly correlated QC

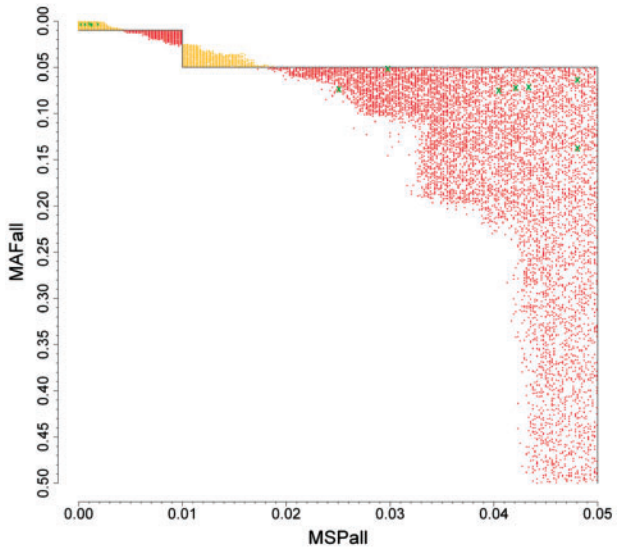

Fig. 4. Characteristics of SNPs with disagreeing classification results between the proposed filter and WTCCC expert filter in CAD. In the graph, each dot represents an SNP; the yellow dots indicate SNPs labeled as 'good' in the proposed filter but 'bad' in the expert filter, and the red dots indicate the opposite scenario. The green '×'s indicate FPs. There are no TPs in the disagreement regions because both filters classified all TPs as 'good' SNPs.

variables, and use DBSCAN to identify the borders of good-SNP clusters that have arbitrary shapes. The boundary of the good-SNP cluster can be translated directly to meaningful thresholds for the original QC variables. The proposed algorithm retains the rationale of the expert filter to identify QC outliers, avoids arbitrary decisions on cutoff values and gives contingent MSP thresholds for different MAF values. The data applications show that with the same or fewer SNPs discarded due to bad quality, the proposed algorithm has comparable or better performance than the expert filter for all diseases.

The underlying rationale of our algorithm is that the majority of genotyped markers have sufficient genotyping quality, and hence low-quality SNPs can be treated as outliers and be identified by looking for SNPs with distinct QC features. To facilitate the implementation of the idea, we use PCA on the original QC variables. PCA consolidates the information from the many correlated QC variables, and projects SNPs onto a two-dimensional PC plane where good SNPs clump together into a corner of desirable QC values, whereas bad SNPs fan out in all directions. It is expected that the PC biplot may differ from one study to another: in our exploration, we have seen different patterns in the biplots for WTCCC datasets and for the MDD dataset. Yet all biplots have good SNPs lumped into a corner that corresponds to good QC features.

We wish to point out that when using the proposed algorithm, it is important to monitor the features of the retained good SNPs and tune the neighborhood radius r to safeguard the basic QC criteria. This is because the thresholds for outliers are determined relative to the majority of the data. The tuning becomes particularly crucial if a big proportion of data points are of low quality. For example, in the MDD dataset, there were about 11.4% of SNPs with MSP $> 10\%$, and our algorithm with the initial change-point r kept SNPs with MSP up to almost 38%. Tuning of r was thus continued until the maximum MSP dropped to $< 10\%$. In practice, the smaller r is, the more stringent the QC criteria for 'good' SNPs will be, as a smaller r makes it harder to form a cluster in DBSCAN. We suggest starting with a value of r determined by fitting a change-point model to the sorted fourth nearest distances, and then further to adjust r until the specific goal is reached, so as to assure the λ value, the maximum MSP, or the percentage of SNPs removed within reasonable ranges. In our explorations, we found that the change-point r often suggests a reasonable value (e.g. in CAD, CD, HT, T1D and T2D) or is at least a good upper bound (e.g. in BD, RA and MDD, judging by the resulting λ values or the retained maximum MSP). Given the change-point r, one can reduce its value if a more stringent filter is needed, and increase its value if one wishes to remove only extreme outlier SNPs.

When selecting which QC variables to include in the algorithm, we intend to avoid using MAFcs and MAFcn because they may obfuscate the true associations. For the rest of the QC variables, it is possible to make other choices for inclusion/exclusion, e.g. to exclude MSPall from the proposed QC variable set (i.e. to use five variables), or to include MSPall/MAFall to the proposed QC variable set (i.e. to use seven variables). While we expected that the performance would not change much, we carried out sensitivity analyses to evaluate the impact of using different QC variables in the proposed algorithm. The results are given in Supplementary Tables 1, 2 and 3. For comparability, we tuned the r values so that each analysis removed a similar proportion of SNPs to the original 6-variable analysis. As expected, the 7- and 5-variable analyses performed very similarly to the proposed 6-variable analyses, indicating the robustness of the proposed filter to the different choices of QC variables.

ACKNOWLEDGEMENTS

This study makes use of data generated by the WTCCC and the GAIN major depression disorder study (GAIN-MDD). For WTCCC, a full list of the investigators who contributed to the generation of the data is available from http://www.wtccc.org.uk. Funding for the WTCCC project was provided by the Wellcome Trust under award 076113. For GAIN-MDD, the genotyping of samples was provided through the Genetic Association Information Network (GAIN), Foundation for NIH. The dataset used for the analyses described in this manuscript were obtained from the database of Genotypes and Phenotypes (dbGaP) found at http://www.ncbi.nlm.nih.gov/gap through dbGaP accession number phs000020.v2.p1. Samples and associated phenotype data for the GAIN-MDD were provided by Drs Patrick Sullivan, Dorret Boomsma, Brenda Penninx, Eco de Geus, Gonneke Willemsen and Witte Hoogendijk. The authors thank Drs Steffen Heber, John Pierre Mertz and Chris Smith for their very helpful discussions on the work.

Funding: National Institutes of Health (R01 MH084022).

Conflict of Interest: none declared.

REFERENCES

Anney,R. *et al*. (2008) Non-random error in genotype calling procedures: implications for family-based and case-control genome-wide association studies. *Am. J. Med. Genet.*, **147B**, 1379–1386.

Barrett,J.C. *et al*. (2008) Genome-wide association defines more than 30 distinct susceptibility loci for Crohn's disease. *Nat. Genet.*, **40**, 955–962.

Clayton,D.G. *et al*. (2005) Population structure, differential bias and genomic control in a large-scale, case-control association study. *Nat. Genet.*, **37**, 1243–1246.

Cox,D.G. and Kraft,P. (2006) Quantification of the power of Hardy-Weinberg equilibrium testing to detect genotyping error. *Hum. Hered.*, **61**, 10–14.

Devlin,B. and Roeder,K. (1999) Genomic control for association studies. *Biometrics*, **55**, 997–1004.

Donnelly,P. (2008) Progress and challenges in genome-wide association studies in humans. *Nature*, **456**, 728–731.

Ester,M. *et al*. (1996) A density-based algorithm for discovering clusters in large spatial databases with noise. In *Proceedings of the 2nd International Conference on Knowledge Discovery and Data Mining*, AAAI Press, Portland, OR, pp. 226–231.

Fardo,D.W. *et al*. (2009) On quality control measures in genome-wide association studies: a test to assess the genotyping quality of individual probands in family-based association studies and an application to the HapMap data. *PLoS Genet.*, **5**, e1000572.

GAIN Collaborative Research Group (2007) New models of collaboration in genome-wide association studies: the Genetic Association Information Network. *Nat. Genet.*, **39**, 1045–1051.

Gordon,D. *et al*. (2001) A transmission/disequilibrium test that allows for genotyping errors in the analysis of single-nucleotide polymorphism data. *Am. J. Hum. Genet.*, **69**, 371–380.

Gordon,D. and Ott,J. (2001) Assessment and management of single nucleotide polymorphism genotype errors in genetic association analysis. *Pac. Symp. Biocomput.*, **6**, 18–29.

Hao,K. and Wang,X. (2004) Incorporating individual error rate into association test of unmatched case-control design. *Hum. Hered.*, **58**, 154–163.

Hindorff,L.A. *et al*. (2009) Potential etiologic and functional implications of genome-wide association loci for human diseases and traits. *Proc. Natl Acad. Sci. USA*, **106**, 9362–9367.

Hugot,J.P. *et al*. (2001) Association of NOD2 leucine-rich repeat variants with susceptibility to Crohn's disease. *Nature*, **411**, 599–603.

McCarthy,M.I. (2004) Progress in defining the molecular basis of type 2 diabetes mellitus through susceptibility-gene identification. *Hum. Mol. Genet.*, **13**, R33–R41.

Miyagawa,T. *et al*. (2008) Appropriate data cleaning methods for genome-wide association study. *J. Hum. Genet.*, **53**, 886–893.

Neale,B.M. and Purcell,S. (2008) The positives, protocols, and perils of genome-wide association. *Am. J. Med. Genet.*, **147B**, 1288–1294.

Pettersson,F. *et al*. (2008) Goldsurfer2 (Gs2): a comprehensive tool for the analysis and visualization of genome wide association studies. *BMC Bioinformatics*, **9**, 1–11.

Plagnol,V. *et al*. (2007) A method to address differential bias in genotyping in large-scale association studies. *PLoS Genet.*, **3**, 0759–0767.

Pompanon,F. *et al*. (2005) Genotyping errors: causes, consequences and solutions. *Nat. Rev. Genet.*, **6**, 847–859.

Purcell,S. *et al*. (2007) PLINK: a toolset for whole-genome association and population-based linkage analysis. *Am. J. Hum. Genet.*, **81**, 559–875.

Rice,K.M. and Holmans,P. (2003) Allowing for genotyping error in analysis of unmatched case-control studies. *Ann. Hum. Genet.*, **67**, 165–174.

Sladek,R. *et al*. (2007) A genome-wide association study identifies novel risk loci for type 2 diabetes. *Nature*, **445**, 881–885.

Sullivan,P.F. *et al*. (2009) Genomewide association for major depressive disorder: a possible role for the presynaptic protein piccolo. *Mol. Psychiatry*, **14**, 359–375.

Teo,Y.Y. *et al*. (2007) On the usage of HWE for identifying genotyping errors. *Ann. Hum. Genet.*, **71**, 701–703.

Teo,Y.Y. (2008) Common statistical issues in genome-wide association studies: a review on power, data quality control, genotype calling and population structure. *Curr. Opin. Lipidol.*, **19**, 133–143.

Teo,Y.Y. *et al*. (2008) Perturbation analysis: a simple method for filtering SNPs with erroneous genotyping in genome-wide association studies. *Ann. Hum. Genet.*, **72**, 368–374.

Unoki,H. *et al*. (2008) SNPs in KCNQ1 are associated with susceptibility to type 2 diabetes in East Asian and European populations. *Nat. Genet.*, **40**, 1098–1102.

Wellcome Trust Case Control Consortium (2007) Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls. *Nature*, **447**, 661–678.

Zeggini,E. *et al*. (2008) Meta-analysis of genome-wide association data and large-scale replication identifies additional susceptibility loci for type 2 diabetes. *Nat. Genet.*, **40**, 638–645.

Ziegler,A. *et al*. (2008) Biostatistical aspects of genome-wide association studies. *Biom. J.*, **50**, 8–28.

APPENDIX A

A.1 THE CHANGE-POINT MODEL FOR ESTIMATING *r*

A typical sorted K-th NN graph is shown in Figure A1a. Given the wide range of the distance, it would be more stable to fit the change-point model to the log transformation of the distance, as shown in Figure A1b. After the log transformation, there are two change points, and our focus is on the right one. The change point model that we consider uses two linear lines to approximate the data points around the change point (Fig. A1c). Let y be the log of the fourth NN distance, x be the (distance-sorted) SNP ID, and x^* be the change point on the X-axis. The two linear models are $y_i = \alpha_0 + \beta_0 x_i + e_i$ for $x_i < x^*$ and $y_i = \alpha_1 + \beta_1 x_i + e_i$ for $x_i > x^*$, where $e_i \sim N(0, \sigma^2)$ and $\alpha_0 + \beta_0 x^* = \alpha_1 + \beta_1 x^*$ (or equivalently $\alpha_1 = \alpha_0 + (\beta_0 - \beta_1)x^*$). A normal likelihood is then specified and optimized to obtain the maximum likelihood estimates $\hat{\alpha}_0, \hat{\beta}_0, \hat{\beta}_1, \hat{\sigma}$ and $\hat{x}^*$. Then the change-point r value is the distance value on the Y-axis corresponding to $\hat{x}^*$.

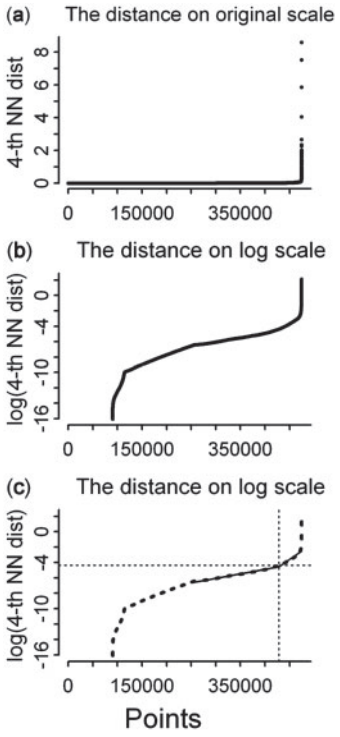

Fig. A1. The sorted fourth NN distance for CAD. (**a**) The distance on the original scale. (**b**) The distance on the log scale. (**c**) The distance on the log scale (dashed line) superimposed with the fit from the change-point model (solid line) as described in the Appendix A1. The change-point r is indicated by the dotted horizontal line.